The Journal of Safety Research is pleased to publish in this special issue the proceedings of several papers presented at the 4th International Conference on Road Safety and Simulation convened at Roma Tre University in Rome, Italy, October 2013. This conference serves as an interdisciplinary forum for the exchange of ideas, methodologies, research, and applications aimed at improving road safety globally.

Conference proceedings provide the opportunity for research in its formative stages to be shared, allowing our readers to gain early insights in the type of work currently being conducted and for the researchers to receive valuable feedback to help inform ongoing activities. This conference in particular offers an array of research topics not often covered by this journal from researchers practicing in over 11 countries. As is common with publishing conference proceedings, the papers published in this issue did not go through the normal JSR review process. Each paper included in this issue did meet the Road Safety and Simulation conference review requirements. They reflect varying degrees of scientific rigor, methodological design, and groundbreaking application.

The proceedings published in this special issue of JSR draw from the following road safety research sectors represented at the conference: driving simulation, crash causality, naturalistic driving, and new research methods.

It is our hope that the publication of these important proceedings will stimulate vigorous dialogue, rigorous research, and continuing innovative initiatives and applications, leading, ultimately, to fewer traffic fatalities, injuries, and crashes.

Thomas W. Planek
Editor-in-Chief

Sergey Sinelnikov
Associate Editor

Jonathan Thomas
Associate Editor

Kenneth Kolosh
Associate Editor

Kathleen Porretta
Managing Editor

18 February 2014
Assessing the risk of secondary crashes on highways

Hong Yang a,b, Kaan Ozbay b, Kun Xie c

a Department of Civil and Urban Engineering, Center for Urban Science and Progress (CUSP), New York University (NYU), One MetroTech Center, 19th Floor, Office 1919N, Brooklyn, NY 11201, USA
b Department of Civil and Urban Engineering, Center for Urban Science and Progress (CUSP), New York University (NYU), One MetroTech Center, 19th Floor, Brooklyn, NY 11201, USA
c Department of Civil and Urban Engineering, New York University (NYU), One MetroTech Center, 19th Floor, Office 1919N, Brooklyn, NY 11201, USA

ARTICLE INFO

Article history:
Received 4 November 2013
Accepted 13 March 2014
Available online 24 April 2014

Keywords:
Secondary crash
Crash risk
Highway
Traffic safety
Rare event logistic regression

ABSTRACT

Introduction: The occurrence of “secondary crashes” is one of the critical yet understudied highway safety issues. Induced by the primary crashes, the occurrence of secondary crashes does not only increase traffic delays but also the risk of inducing additional incidents. Many highway agencies are highly interested in the implementation of safety countermeasures to reduce this type of crashes. However, due to the limited understanding of the key contributing factors, they face a great challenge for determining the most appropriate countermeasures. Method: To bridge this gap, this study makes important contributions to the existing literature of secondary incidents by developing a novel methodology to assess the risk of having secondary crashes on highways. The proposed methodology consists of two major components, namely: (a) accurate identification of secondary crashes and (b) statistically robust assessment of causal effects of contributing factors. The first component is concerned with the development of an improved identification approach for secondary accidents that relies on the rich traffic information obtained from traffic sensors. The second component of the proposed methodology is aimed at understanding the key mechanisms that are hypothesized to cause secondary crashes through the use of a modified logistic regression model that can efficiently deal with relatively rare events such as secondary incidents. The feasibility and improved performance of using the proposed methodology are tested using real-world crash and traffic flow data. Results: The risk of inducing secondary crashes after the occurrence of individual primary crashes under different circumstances is studied by employing the estimated regression model. Marginal effect of each factor on the risk of secondary crashes is also quantified and important contributing factors are highlighted and discussed. Practical applications: Massive sensor data can be used to support the identification of secondary crashes. The occurrence mechanism of these secondary crashes can be investigated by the proposed model. Understanding the mechanism helps deploy appropriate countermeasures to mitigate or prevent the secondary crashes.

© 2014 National Safety Council and Elsevier Ltd. All rights reserved.

1. Introduction

The presence of traffic incidents such as crashes, disabled vehicles, and debris on the road is a major contributing factor that reduces capacity and service quality of transportation systems. These incidents account for approximately one-fourth of all traffic delays (FHWA, 2007). More severely, these incidents will induce secondary crashes that put motorists and responder lives at risk. As shown by Tedesco, Alexiadis, Loudon, Margiotta, and Skinner (1994), the crash risk will increase more than six times in case a prior crash occurred. Similarly, an additional minute increase in the clearance time increases the likelihood of secondary crashes by 2.8% (Karlaftis, Latoski, Richards, Nadine, & Sinha, 1999). It is estimated that secondary crashes alone account for 20% of all crashes and 18% of all fatalities on freeways (O’Laughlin & Smith, 2002; Owens et al., 2010). In addition, secondary crashes induce additional traffic congestions and delays to road users. Therefore, the prevention of secondary crash has been placed in a high priority in traffic incident management (O’Laughlin & Smith, 2002).

In order to develop appropriate countermeasures to mitigate secondary crashes, it is necessary to understand the mechanism of its occurrence. Despite earlier efforts on investigating these crashes, there are still very limited studies focusing on mining the casual relationship between secondary crashes and possible explanatory variables. Most existing work has focused on identifying secondary crashes (Chou & Miller-Hooks, 2010; Green, Pigman, Walton, & McCormack, 2012; Moore, Giuliano, & Cho, 2004; Raub, 1997a,b; Sun & Chilukuri, 2010; Yang, Barton, & Ozbay, 2013; Yang, Morgul, Barton, & Ozbay, 2014; Zhan, Gan, & Hadi, 2009) and analyzing their corresponding characteristics (Hirunyanitwattana & Mattingly, 2006; Yang, Barton, & Ozbay, 2013a; Zhan, Shen, Hadi, & Gan, 2008; Zhang & Khattak, 2010b). Therefore, the objective of this paper is to examine the mechanism of secondary crash occurrence. The relationship between...
secondary crash occurrence and a number of contributing factors is explored using the identified primary–secondary crash pairs based on an improved identification approach.

2. Literature review

A number of studies have conducted secondary crash analysis. However, the majority of them are focused on the development of the approaches to identify secondary crashes as a consequence of priori crashes (Chilukuri & Sun, 2006; Chou & Miller-Hooks, 2010; Green et al., 2012; Raub, 1997a,b; Sun & Chilukuri, 2005, 2010; Yang et al., 2013; Yang, Morgul, Barton, & Ozbay, 2014s; Zhan et al., 2009). Both static and dynamic identification approaches were proposed and implemented. Another group of studies explores the characteristics of secondary crashes (Hirunyanitwattana & Mattingly, 2006; Kopitch & Saphores, 2011; Vlahogianni, Karlaftis, Golas, & Halkias, 2010; Yang et al., 2013a; Yang, Barton, & Ozbay, 2013b; Zhan et al., 2008; Zhang & Khattak, 2010a, 2011).

Only a limited number of studies examined the factors that may affect the occurrence of secondary crashes and/or the corresponding countermeasures. For instance, Karlaftis et al. (1999) and Latoski, Pal, and Sinha (1999) developed a logistic regression model to examine the likelihood of secondary crashes based on the characteristics of the primary incident. Similar modeling approaches were also used by Zhan et al. (2008), Zhan et al. (2009), Kopitch and Saphores (2011), and Khattak, Wang, and Zhang (2012). These models were also used to support the development of traffic incident management tools. For example, they were used to evaluate effects of highway service patrol program (Karlaftis et al., 1999; Latoski et al., 1999) and changeable message signs (Kopitch & Saphores, 2011) on reducing secondary crashes. Zhang and Khattak (2010b) used ordered logit models to explore the probability of multiple secondary crashes. It was found that the number of vehicles involved and the number of lane closures had a different impact on the occurrence of primary–secondary crash pairs and primary–multiple-secondary crash pairs.

Other than modeling the causal-relationship between prior crashes and secondary crashes, Zhan et al. (2008, 2009) also used logistic regression to identify the factors that affect the severity of secondary crashes. They found that lane blockage duration and visibility significantly affect secondary crash severity. Khattak, Wang, and Zhang (2010) further predicted the frequency of secondary incidents from a macroscopic level. Poisson, zero-inflated Poisson, and negative binomial regression models were estimated. They found that factors including roadway length, traffic volume, number of on-ramps, curve level, number of lanes, congestion level, truck volumes, and roadway location affected the frequency of secondary incidents.

Despite the efforts on modeling the occurrence of secondary crashes, there is still no clear and consistent understanding of the mechanism of the secondary crash occurrence. This might be attributed to two major issues: (a) the lack of high-quality incident data and (b) the lack of a consistent approach to accurately identify secondary crashes. The first issue mainly limits our capability of fully understanding the characteristics of the crashes. The two issues together also lead to biased classification of secondary crashes. Consequently, any analysis or model based on the misclassified secondary crashes will lead to biased findings. These unreliable findings in turn lead to unreliable decisions on selecting appropriate incident management countermeasures. Therefore, a thorough examination of secondary crash occurrence based on reliable identification results is needed.

3. Methodology

The risk of having secondary crashes is of great interest in this study. Hereinafter the risk is defined as the probability of a prior crash that induces one or more secondary crashes. Factors that may affect the secondary crash risk are investigated based on the improved identification outcomes.

3.1. Identifying secondary crashes

In general, secondary crashes are defined as crashes that occurred within the spatial and temporal impact ranges of a prior crash. Most of the existing studies still rely on static approach to identify secondary crashes (Green et al., 2012; Hirunyanitwattana & Mattingly, 2006; Karlaftis et al., 1999; Moore et al., 2004; Raub, 1997a,b). These approaches are not adequate as they need the subjective selection of fixed spatio-temporal thresholds to describe the impact range of the prior crashes. Similarly, those approaches based on queuing models also lack the accuracy to fully capture secondary crashes (Chou & Miller-Hooks, 2010; Haghani, Ilescu, Hamedi, & Yong, 2006; Vlahogianni et al., 2010; Zhang & Khattak, 2011). Yang, Barton, and Ozbay (2013) recently developed an improved identification methodology to accurately account for the dynamic characteristics of the spatio-temporal impact of primary incidents under the prevailing traffic conditions. The proposed approach intends to identify the impact range of a prior crash using traffic information mined from archived sensor data, and to detect secondary crashes within the impact range. The detailed description of the methodology can be found in Yang, Barton, and Ozbay (2013). The general procedure is summarized as follows.

Step 1 Develop speed contour plot (SCP) over time and space: Let’s use Fig. 1(a), which is an example extracted from the integrated traffic sensor data and crash database described in the next section. Each cell in the figure represents a speed measurement $V(t,s)$ from a loop detector $s$ along the freeway at the $t$th time interval, $s = 1,2,...,S$ and $t = 1,2,...,T$. Fig. 1(a) represents a slice of a typical SCP on a freeway. A clear queue formation is observed soon after crash A.

Step 2 Develop a representative speed contour plot (RSCP): The RSCP generally represents daily normal traffic conditions on a freeway when no incident exists. Such RSCP can be built based upon representative speed measurements. The $p$th percentile speed $V_p(t,s)$ of the historical speed measurements on the $i$th day of the week is used as the representative speed measured at detector $s$ and time period $t$, where $i = 1,2,...,7$ analogous to the day of the week from Monday to Sunday. Fig. 1(b) shows an example of RSCP that represents a normal Wednesday traffic condition on a freeway based on data collected every 5 min during all other Wednesdays in 2011.

Step 3 Construct a binary speed contour plot (BSCP): Compare $V(t,s)$ of SCP in step 1 with the corresponding $V(t,s)$ in RSCP generated in step 2. If $V(t,s) < ω × V_p(t,s)$, the speed measurement $V(t,s)$ in the original SCP is converted into $V(t,s) = 1$. Otherwise, it is denoted as $V(t,s) = 0$. This constraint defines an abnormal traffic condition if the measured speed is below the threshold $ω × V_p(t,s)$, where $ω$ is a user defined weighting factor between 0 and 1. It assumed that a $(1 − ω) × 100\%$ reduction in the representative (normal) speed indicates the occurrence of non-recurrent congestion. A small weight factor suggests an aggressive threshold to define non-recurrent congestion whereas a large one implies a conservative threshold. A reasonable weight factor should be determined based on highway operator’s classification of congested and non-congested conditions. To reflect the consistency of speed measurements in short time periods, the speed measurement at the $t$th time period is changed to $V(t,s) = 1$ if
\[ \{ V(t-1,s) = 1, V(t,s) = 0, \text{ and } V(t+1,s) = 1 \} \]. After conversion, the original SCP will be represented by a BSCP. Fig. 1(c) shows an example of converting the original SCP into a BSCP based on the RSCP of 50th percentile historical speed. In BSCP, a red cell indicates that \( V(t,s) = 1 \) and a green cell means \( V(t,s) = 0 \). Each cluster of red cells is used to visualize the congested area.

Step 4: Detect secondary crashes in the impact range in BSCP: Let’s use Fig. 1(d) as an example. We need to detect whether crash B and crash C are secondary crashes in relation to the first crash.

Fig. 1. Illustration of identifying secondary crashes using sensor data. Yang et al. (2013a).
A. The line (e.g., lines AB and AC) between each pair of potential primary–secondary crashes is estimated based on their coordinates. If the impact range contains the line, then it is suggested that this pair of crashes is primary and secondary ones. Otherwise, they are independent crashes. Readers may refer to Yang, Bartin, and Ozbay (2013) for the detailed description of the algorithm for detecting whether the line is enclosed by the impact range. The essential idea of the algorithm is to check the coordinates of some critical points on the line in relation to the impact range. If the binary speed measurements for cells that contain all the sampled points are one, then it suggests that the line is contained within the impact range and the corresponding crashes are primary–secondary crashes (e.g., crashes A and B). Otherwise, they are assumed to be independent crashes (i.e., crashes A and C).

The aforementioned steps are used to identify the secondary crashes when a prior crash causes observable traffic impact through the sensor data. Some minor crashes that may not cause significant queues can still lead to secondary crashes as a result of distracted drivers. In this case, we assume that secondary crashes only occur within half an hour and half a mile upstream of prior crashes. Crashes that occur in the opposite direction within 1 h and one mile upstream of prior incidents are also assumed to be the result of rubbernecking. In addition to these criteria, a vehicle crash in the opposite direction is identified as secondary only if it occurred in a queue.

### 3.2. Modeling the risk of secondary crashes

Denote \( y \) as the occurrence of a secondary crash. It can be described as a binary outcome. Let \( y = 1 \) represent that a secondary crash is induced by a primary crash whereas \( y = 0 \) indicates no secondary crash occurred. A binary logistic regression model can be developed to examine the influence of factors that may affect the secondary crash risk of a prior crash.

Let us define \( \pi(x) \) as the probability of a secondary crash induced by a prior crash and \( 1 - \pi(x) \) as the probability of having no secondary crash. The binary logistic regression model identifies the relationship between the log odds of the binary outcome and various risk factors. It can be formulated as following Eq. (1):

\[
\begin{align*}
\log[\pi(x)] &= \log\left(\frac{\pi(x)}{1-\pi(x)}\right) = \alpha + X^T \beta.
\end{align*}
\]

Based on the above equation, the probability that a prior crash inducing a secondary crash can be described by the logistic distribution shown in following Eq. (2):

\[
P(y = 1|X) = \pi(x) = \frac{\exp(\alpha + X^T \beta)}{1 + \exp(\alpha + X^T \beta)}
\]

where:

- \( \pi(x) \) the conditional probability of the form \( P(y = 1|X) \)
- \( X \) the vector of contributing factors that could be continuous or discrete
- \( \beta \) the corresponding vector of the coefficients
- \( \alpha \) an intercept parameter.

The maximum-likelihood estimation (MLE) technique can be used to estimate the regression model’s parameters. The likelihood function is constructed as shown in Eq. (3). By maximizing the log likelihood expression shown in Eq. (4), the best estimates of parameters \( \alpha \) and \( \beta \) can be obtained accordingly.

\[
\begin{align*}
l(\beta) &= \prod_{i=1}^{n} \left( \pi(x_i)^{y_i} \left[1-\pi(x_i)\right]^{(1-y_i)} \right)
\end{align*}
\]

\[
\begin{align*}
\text{LL}(\beta) &= \ln[l(\beta)] = \sum_{i=1}^{n} \left[ y_i \ln[\pi(x_i)] + (1-y_i) \ln[1-\pi(x_i)] \right]
\end{align*}
\]

where:

- \( y_i \) the observed outcome of the \( i \)th prior crash, with the value of either 0 or 1 only
- \( n \) total number of prior crashes.

The overall goodness-of-fit of the logistic regression model is tested by using the likelihood ratio test. The significance of individual contributing factors within the model is examined using the Wald z statistic. Moreover, the influence of \( j \)th factor on the occurrence of a secondary crash can be revealed by the odd ratio (OR), which is defined as Eq. (5):

\[
\begin{align*}
\text{OR} &= \exp(\beta_j)
\end{align*}
\]

where:

- \( \beta_j \) the coefficient of the \( j \)th factor.

OR measures the ratio of the predicted odds for a one-unit increase in a continuous variable \( x_j \) or the presence of a discrete variable \( x_j \) when other variables in the model are held constant. An OR greater than one suggests that the \( j \)th factor increases the likelihood of secondary crash occurrence when a prior crash occurs and vice versa.

Although the binary logistic regression models have been frequently used in transportation safety studies, these models can yield biased estimates when sample data exhibits class imbalance (Buehler, 2012). In this study, secondary crashes induced by prior crashes can be an example of class imbalance because most crashes do not induce secondary crashes. More importantly, there are many crashes that did not induce a secondary crash (\( y = 0 \)) whereas only a small number of them had induced secondary crashes (\( y = 1 \)). Thus these secondary crashes can be denoted as rare events. In such cases, the MLE of the binary logistic regression model can result in biased coefficients that underestimate probabilities of the occurrence of secondary crashes (King & Zeng, 2001a,b). To avoid these estimation issues, King and Zeng (2001b) have developed the rare event logistic regression for the cases that include three types of corrections for the ordinary logistic regression. First, a case-control sampling design based on endogenous stratified sampling is recommended. It consists of taking all events (\( y = 1 \)) and a random selection of the non-events (\( y = 0 \)). The proportion of events to non-events is usually set around 1:10 (Beguería, 2006; Ramalho, 2002). The second correction step is called prior correction. It intends to avoid dependent variable selection that may introduce sampling bias on the logistic coefficients (King & Zeng, 2001a,b). The intercept \( \alpha \) is corrected using the following equation.

\[
\begin{align*}
\alpha &= \alpha - \ln \left( \frac{1-\tau}{\tau} \times \frac{\bar{y}}{1-\bar{y}} \right)
\end{align*}
\]

where:

- \( \alpha \) the corrected intercept
- \( \alpha \) the uncorrected intercept
- \( \tau \) the actual true fraction of 1s (events) in the population
- \( \bar{y} \) the observed fraction of 1s (events) in the sample.

The third correction step is to adjust the underestimation of the probabilities when using the corrected intercept \( \alpha \). A correction factor
$C_i$ is incorporated to the estimated probability $\hat{p}_i$ to obtain the corrected probability $\tilde{p}_i$.

$$\tilde{p}_i = \hat{p}_i + C_i,$$

The correction factor $C_i$ is obtained using the following equation:

$$C_i = (0.5 - \hat{p}_i)\hat{p}_i(1 - \hat{p}_i)XV(\beta)^X'$$

where:

- $\hat{p}_i$ is the event probability estimated using the bias-corrected coefficient $\alpha$.
- $X$ is a $1 \times (n + 1)$ vector of values for each explanatory variable.
- $V(\beta)$ is the variance–covariance matrix.

To implement the rare event logistic regression model, the “relogit” function from the open source R package Zelig (Imai, King, & Lau, 2007) is used.

4. Data sources and variables

To investigate the mechanism of secondary crash occurrence, a 27-mile section of the New Jersey Turnpike (NJTPK) between interchanges 5 and 9 was used as the case study. This is one of the busiest tolled freeways in the United States crossing New Jersey from south to north. Fig. 2 illustrates the site map. The section has 25 remote traffic microwave sensors (RTMS) placed approximately at every one mile on the mainline, which provide rich traffic data for testing the proposed secondary crash identification approach. The speeds aggregated in 5-minute intervals were extracted to develop the speed contour plots using the proposed identification method.

Traffic crash data for this section in 2011 were extracted from the online crash database of the New Jersey Department of Transportation (NJDOT, 2013). These crash data include the detailed crash information such as date, time, location, crash type, number of vehicles involved, vehicle characteristics, and crash severity. After the exclusion of two crashes occurred on ramps and other 24 with unknown directions, the remaining data consist of a total of 1,188 crash records for the studied section (southbound: 575 crashes; northbound: 613 crashes). The secondary crashes among the 1,188 crash records were identified using the aforementioned identification approach. Since the crash data do not provide information of the incident duration, a separate dataset that contains durations was used. The incident records in the year 2011 were linked with the crash records and the duration information of each crash was then extracted.

Based on information collected from these separate data sources, relationship between the variables shown in Table 1 and the occurrence of secondary crashes was examined.

5. Results and discussion

The proposed approach for the identification of secondary crashes was implemented using the above integrated dataset and 71 primary crashes were found to induce 100 secondary crashes (note: one primary crash may cause multiple secondary crashes). Yang, Bartin, and Ozbay (2013) compared this novel approach with the traditional identification approaches (Moore et al., 2004; Raub, 1997a). It has been shown that the proposed method for identifying secondary incidents reduced the incorrect classifications and captured additional secondary crashes missed by the traditional static threshold methods. Detailed characteristics of these secondary crashes were examined in Yang et al. (2013b). To model the occurrence risk of these secondary crashes, choice-based sampling design based on the endogenous stratified sampling was used. Specifically, a proportion of 1:10 for the ratio of events (crashes that induce secondary crashes) to non-events (crashes that induce no

Table 1
Variable names, definitions, and descriptive statistics.

<table>
<thead>
<tr>
<th>Variables</th>
<th>Type</th>
<th>Description</th>
<th>Mean</th>
<th>Sd.</th>
<th>Min.</th>
<th>Max.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Occurrence</td>
<td>Indicator</td>
<td>Secondary crash occur = 1; otherwise = 0</td>
<td>0.07</td>
<td>0.25</td>
<td>0.00</td>
<td>1.00</td>
</tr>
<tr>
<td>Time period</td>
<td>Categorical</td>
<td>Nighttime off-peak period (19 pm to 7 am) = 1; Daytime off-peak period (9 am to 17 pm) = 2; Daytime peak period (7 to 9 am or 17 to 19 pm) = 3</td>
<td>1.84</td>
<td>0.66</td>
<td>1.00</td>
<td>3.00</td>
</tr>
<tr>
<td>Rear end</td>
<td>Indicator</td>
<td>Prior crash is rear end crash = 1; otherwise = 0;</td>
<td>0.44</td>
<td>0.50</td>
<td>0.00</td>
<td>1.00</td>
</tr>
<tr>
<td>Severity</td>
<td>Indicator</td>
<td>Prior crash is injury or fatal crash = 1; otherwise = 0</td>
<td>0.20</td>
<td>0.40</td>
<td>0.00</td>
<td>1.00</td>
</tr>
<tr>
<td>Duration</td>
<td>Numerical</td>
<td>Duration of prior crash (minute)</td>
<td>50.9</td>
<td>36.0</td>
<td>10.0</td>
<td>367.0</td>
</tr>
<tr>
<td>Work zone</td>
<td>Indicator</td>
<td>Prior crash occurs at work zone = 1; otherwise = 0</td>
<td>0.39</td>
<td>0.49</td>
<td>0.00</td>
<td>1.00</td>
</tr>
<tr>
<td>Weekend</td>
<td>Indicator</td>
<td>Prior crash occurs on weekend = 1; otherwise = 0</td>
<td>0.30</td>
<td>0.46</td>
<td>0.00</td>
<td>1.00</td>
</tr>
<tr>
<td>Winter</td>
<td>Indicator</td>
<td>Prior crash occurs in winter = 1; otherwise = 0</td>
<td>0.17</td>
<td>0.37</td>
<td>0.00</td>
<td>1.00</td>
</tr>
<tr>
<td>Lane closure</td>
<td>Indicator</td>
<td>More than one lane closure = 1; otherwise = 0</td>
<td>0.02</td>
<td>0.14</td>
<td>0.00</td>
<td>1.00</td>
</tr>
<tr>
<td>Truck involved</td>
<td>Indicator</td>
<td>Prior crash involves truck = 1; otherwise = 0</td>
<td>0.30</td>
<td>0.46</td>
<td>0.00</td>
<td>1.00</td>
</tr>
</tbody>
</table>
secondary crash) was used according to the recommendations identified from the literature (Beguería, 2006; Ramalho, 2002).

The estimation results are presented in Table 2. Only the variables that are statistically significant, namely, at the significance level of 0.1, are included in the rare event logistic regression model. The results show that compared with the night-time off-peak periods, crashes that occur during the day time off-peak hours have higher likelihood of inducing secondary crashes. The corresponding odd ratio is 4.906. Similarly, crashes that occur during the day-time peak hours have the highest likelihood of inducing secondary crashes, with odd ratio of 6.704. These quantified impacts of the occurrence time of prior crashes on the risk of secondary crashes are consistent with previous reported results (Khattak et al., 2009; Zhan et al., 2008, 2009). Interestingly, compared with other types of crashes, rear end crashes tend to increase the risk of having secondary crashes. Specifically, if the prior crash is a rear end crash, the probability of inducing a secondary crash is 71.3% higher than that of other types of crashes.

As expected, incident duration is positively associated with the occurrence of secondary crashes. Longer duration leads to higher likelihood of secondary crash occurrence. According to the odd ratio, each additional minute increase in the duration increases the likelihood of a secondary crash by 1.2%. This finding is slightly different from the results reported in literature (Karlaftis et al., 1999; Latoski et al., 1999), where 1 min increase in the clearance time increases the likelihood of secondary crash by 1.8 to 3.2%. The difference should be attributed to the approaches used to identify secondary crashes and to model the secondary crash occurrence. As found in previous studies (Zhan et al., 2008, 2009), the number of lanes closed also is found to have a statistically significant effect on the occurrence of secondary crashes. The corresponding odd ratio indicates that if a prior crash caused the closure of more than one lane, then the risk of having secondary crashes will be 4.811 times higher than those minor crashes that cause no more than one lane closure. The negative coefficient of “winter” covariate indicates that crashes that occurred in winter were less likely to cause secondary crashes compared with other seasons. This finding is consistent with the results reported in Karlaftis et al. (1999). In their study, it was argued that drivers are inherently more careful in winter and drive at lower speeds, which reduces the likelihood of a secondary crash. Another potential reason might be attributed to the relative lower level of demand in winter season. A closer look at the relationship between traffic variation and season would be helpful.

Unlike the significant variables reported in Karlaftis et al. (1999) and Latoski et al. (1999), day of the week and truck involvement did not significantly affect the occurrence of secondary crashes in this study. Moreover, variables including crash severity and work zone shoulder closure also had no significant impact on the likelihood of secondary crashes.

6. Conclusions

The main goal of this study is to analyze the factors affecting the occurrence risk of secondary crashes given the occurrence of a prior crash on a highway. To achieve this goal, two major tasks have been performed: (a) improving identification of secondary crashes from the crash database and (b) enhanced modeling of the occurrence risk of the identified secondary crashes. An improved identification approach that makes use of combined large sensor and crash datasets was implemented to identify most likely secondary crashes. To identify the impact of different variables on the occurrence of a secondary crash, a “rare event logistic” regression model has been specified and estimated. The rare event logistic regression model is used to address the fact that the secondary crashes are relatively rare events compared with regular crashes. A numerical study using data from New Jersey has been conducted. First, secondary crashes were identified using the developed novel identification approach. Then the occurrence risk of the identified secondary crashes was modeled using the proposed rare event logistic regression. Explanatory variables including time periods, crash type, crash duration, number of lanes closed, and season were found to significantly affect the likelihood of secondary crash occurrence. In addition, variables such as crash severity, initial crash occurred at a work site, truck involvement, and day of the week were not found to significantly contribute to the likelihood of secondary crashes. To reduce the risk of secondary crashes, programs that promote fast clearance of incidents should be implemented and vehicles involved in crashes should be removed from the travel lanes as soon as possible. An additional minute increase in the incident duration will increase the likelihood of secondary crash occurrence by 1.2%. Therefore, it is important to allocate more resources during the peak periods to reduce incident response and process times to reduce number of secondary incidents.

Acknowledgments

The authors thank the New Jersey Turnpike Authority and the New Jersey Department of Transportation for proving the necessary data. The contents of this paper reflect views of the authors who are responsible for the facts and accuracy of the data presented herein. The contents of the paper do not necessarily reflect the official views or policies of the agencies mentioned in the paper. The author also appreciates the anonymous reviewers from the 4th International Conference on Road Safety and Simulation (RSS2013) for their constructive comments to help improve the earlier version of this paper. The authors thank for the recognition and recommendation from the RSS2013 committees.

References


Table 2 Results of rare event logistic regression modeling.

| Independent variables | Estimate | Std. error | z value | Pr(>|z|) | Odd ratio |
|-----------------------|----------|------------|---------|----------|-----------|
| (Intercept)           | −4.889   | 0.517      | −9.455  | 0.000    | −         |
| Daytime off-peak hours| 1.591    | 0.490      | 3.247   | 0.001    | 4.906     |
| Daytime peak hours    | 1.903    | 0.553      | 3.443   | 0.001    | 6.704     |
| Rear end              | 0.538    | 0.277      | 1.945   | 0.052    | 1.713     |
| Duration              | 0.012    | 0.003      | 3.928   | 0.000    | 1.012     |
| Lane closure          | 1.571    | 0.628      | 2.502   | 0.012    | 4.811     |
| Winter                | −0.835   | 0.468      | −1.783  | 0.075    | 0.434     |


